Behav Genet (2011) 41:625–628
DOI 10.1007/s10519-010-9427-0

BRIEF COMMUNICATION

Genetical Genomic Analysis of Complex Phenotypes Using the PhenoGen Website

Beth Bennett · Laura M. Saba · Cheryl K. Hornbaker · Katerina J. Kechris · Paula Hoffman · Boris Tabakoff

Received: 16 November 2010 / Accepted: 3 December 2010 / Published online: 24 December 2010

Abstract Our laboratory has developed an online interactive resource called PhenoGen (http://phenogen.ucdenver.edu) which provides an archive of brain and other organ gene expression data from a panel of 20 common inbred mouse strains, and three recombinant inbred (RI) panels (two mouse and one rat). DNA microarray data can also be uploaded to the site where numerous analytical tools can be implemented. An important advantage to the archived data is that each array represents data from a single animal and each strain was sampled 4–7 times, providing an estimate of genetic variance (heritability) of individual transcript levels. These panels also allow genetic mapping of expression QTLs. Overlap of eQTLs with phenotypic QTLs provides a powerful approach to candidate gene identification. These methods are briefly described here and we encourage the use of our site for both scientific discovery and as a teaching tool in quantitative genetics.

Keywords Genetical genomics · Gene expression · QTL mapping · eQTL · Recombinant inbred strains

The advent of DNA microarray technology allowed simultaneous, unbiased measurement of levels of most mRNA transcripts. These measures can be used to address many questions in biology, since variations in the expression levels of the mRNA, i.e., amount of mRNA available for translation, have been identified as etiologic factors for behavioral, physiological, or disease phenotypes (Schadt et al. 2005). Differential expression of genes in selected strains, or treated and control tissues, can be used to identify candidate genes influencing a phenotype. However, the set of candidate genes may contain a large proportion of false positives. One method for reducing the number of false positives is to require that the area of the genome that controls the transcription of a candidate gene also controls variation in the phenotype of interest. To do this, transcript levels measured in segregating populations are treated as quantitative traits. Using genetic marker information, chromosomal regions that regulate transcript levels are mapped as expression quantitative trait loci (eQTLs). This approach, "genetical genomics", has been successfully used to construct regulatory gene transcription networks in various organisms (Schadt et al. 2005; Chesler et al. 2005; Lu et al. 2008) and extended to identify candidate genes contributing to complex behavioral phenotypic traits (Saba et al. 2006; Tabakoff et al. 2008). The identification of common QTLs regulating both the phenotype of interest (pQTLs), and the expression of genes whose transcript levels correlate with the phenotype, provides a strong basis for candidate gene identification (Chesler et al. 2005; Hu et al. 2008; Saba et al. 2006; Tabakoff et al. 2008).

The use of microarrays, though attractive, is out of reach for many researchers without access to costly facilities, supplies, and technicians required to process tissue for gene expression and interpret results. We have developed an online, interactive resource (http://phenogen.ucdenver.edu) that enables researchers to mine data from microarray

Edited by John Hewitt.

B. Bennett (✉) · L. M. Saba · C. K. Hornbaker · P. Hoffman · B. Tabakoff
Department of Pharmacology, University of Colorado Denver School of Medicine, PO Box 6511, Mail Stop 8303, Aurora, CO 80045-0511, USA
e-mail: beth.bennett@ucdenver.edu

K. J. Kechris
Department of Biostatistics and Informatics, University of Colorado Denver, Colorado School of Public Health, Aurora, CO 80045, USA

expression experiments and implement the genetical genomic/phenotypic approach described above. Users of the website can access microarray data available on the website (or upload their own microarray data), then use a variety of analytical methods, some of which we describe here for candidate gene searches using their own or published data on quantitative traits in panels of animals for which our site stores genetic and genomic data. We have previously demonstrated the use of the PhenoGen website to identify candidate genes influencing sensitivity to morphine analgesia (Hoffman et al. 2010) and ethanol drinking (Saba et al. 2006; Tabakoff et al. 2008, 2009).

The PhenoGen site allows access to and manipulation of whole brain gene expression data from the BXD recombinant inbred (RI) mouse panel (32 strains and 172 arrays), the HXB/BXH RI rat panel (described by Printz et al. 2003) (22 strains and 139 arrays), and 20 commonly used inbred strains of mice (90 arrays). The site also has data available on transcript expression on mRNA at the exon level for whole brain from the LXS RI mouse panel (describe by Williams et al. 2004) and for heart RNA from the HXB/BXH rat panel (mRNA data from liver, intestine and fat tissue will soon be available).

Several important advantages of the PhenoGen data and associated analytical tools are:

1. each array represents mRNA generated from tissue from a single animal;
2. data on 4–7 biological replicates are available (i.e., one animal's RNA per array), allowing calculation of heritability for each transcript and increased power;
3. datasets from the RI panels and the inbred strain set have been normalized and cleaned of ambiguous and non-unique probes as well as probes containing SNPs (see below), but users can carry out their own cleaning and normalization procedures as well;
4. users can create, normalize, and save their own datasets by selecting from available arrays or using their own uploaded data. By registering on the site, users can create private repositories for saving data and analyses;
5. users can implement various 'filters' (see below) prior to statistical analysis to reduce the multiple testing burden by integrating relevant prior knowledge.

The analysis on the PhenoGen website includes a probe mask that eliminates probes containing known SNPs, and probes that are not targeted to a unique location in their respective genome. This mask has been implemented on the Affymetrix Mouse and Rat Exon arrays and on the Affymetrix Mouse 430 v2 array.

Users have many tools for analyzing raw data, including a choice of normalization procedures, several filtering methods for eliminating data from non-informative or noisy probesets, the ability to choose appropriate statistical methods for data analysis, as well as access to comprehensive annotation and literature search capabilities. The use of these PhenoGen tools is described in more detail by Hoffman et al. (2010), the Demo link on the website, and the User's Manual that can be downloaded from the website.

A typical workflow for a genetical genomic/phenomic approach to identifying candidate genes for a complex quantitative trait is described below. To complete these analyses the user must have phenotypic data from at least 5 of the rat or mouse strains from one of the panels on our website, for correlation, or 15 strains for pQTL mapping. Phenotypes can be uploaded or typed directly into a dialog box. Phenotypic measures can be user-generated or obtained from online or literature sources (see Hoffman et al. 2010).

Candidate genes are identified through a series of filtering steps:

1. Eliminate transcripts not expressed in brain, i.e., probesets with expression values below detection limits;
2. Retain transcripts with evidence for strong genetic heritability of mRNA levels;
3. Retain transcripts whose expression level is controlled within the same area of the genome (eQTL) that controls the variation in the phenotype (pQTL);
4. Retain transcripts whose expression level significantly correlates with the quantitative phenotype (steps 3 and 4 can be performed in reverse order).

The measure of heritability gives insight into the strength of genetic control of transcript expression, i.e., correlation of expression level with phenotype represents a true genetic relationship, rather than an association with environmental variables. Heritability has important implications for the relevance of a transcript/protein for drug targeting or disease characterization.

As described in our previous work, the rationale for using the eQTL/pQTL overlap filter is that if a gene contributes to a complex behavior through variation in its expression levels, areas of the genome that regulate those expression levels (eQTLs) should be represented within areas of the genome that contribute to the regulation of the phenotype. Transcripts with eQTL/pQTL overlap are considered to represent likely candidate genes for the particular phenotype (Hu et al. 2008; Tabakoff et al. 2008). The determination of eQTLs for the panels on the website was described in earlier work (Hu et al. 2008; Tabakoff et al. 2008, 2009), and the eQTL locations (obtained using gene expression data from RI strains), including 95% confidence intervals for eQTLs with empirical p values < 0.1, are available on PhenoGen.

Finally, for quality control purposes, outlier arrays can be easily and confidently identified by comparison to other arrays within a strain. This check has been done for the 'public' datasets of the RI and inbred strain panels, which have been normalized and probe-masked, and can be accessed by a single click on the website.

Because of the multitude of filtering steps involved in the approach described here, the probability of missing a "true" candidate gene is increased. However, the probability of including a "false" candidate gene in the final set of candidate genes is greatly reduced. For example, we examined differential expression in three strain pairs, each of which shows large differences in ethanol consumption (B6 and D2; ILS and ISS; HAP and LAP), and used the method described above to search for candidate genes influencing this phenotype. One probeset passed all filters and showed a significant correlation with ethanol intake, Gnb1 (guanine nucleotide binding protein, β1 subunit; unpubl. data), replicating previous results implicating Gnb1 in drinking (Saba et al. 2006; Tabakoff et al. 2008). A second analysis, using published measurements of sensitivity to morphine-induced analgesia in BXD RI strains (Bergeson et al. 2001) and gene expression data from the same strains identified candidate genes known to affect recycling of opiate receptors and receptor glycosylation, providing a mechanistic explanation for the range in analgesia seen in the BXD (Hoffman et al. 2010).

We have previously described some of the differences between functions available on the PhenoGen site and other sites such as GeneNetwork (Hoffman et al. 2010). Briefly, PhenoGen offers the ability for extensive quality control, user-generated datasets, and a broad choice of statistical analysis. PhenoGen is organized primarily to allow for genetical genomic/phenotypic analysis that includes the determination of bQTL/eQTL overlap as a key aspect of candidate gene identification. However, the site also provides other tools for gene expression analysis, such as promoter analyses (oPOSSUM and MEME), and co-expression/clustering analyses, including graphical tools for visualization of these results.

We continue to improve the PhenoGen website by adding new genetic/genomic data, adding the latest tools for systems genetics approaches, and improving usability. By the spring of 2011, we will have added brain expression data from 64 LXS strains [derived from ILS × ISS crosses (Williams et al. 2004)], heart, liver, gut, and fat exon expression data from 27 HXB/BXH rat strains, and new genotype information for both of these panels. We also plan to expand our functionality to include analysis of high throughput RNA sequencing data and to implement popular biologic network analysis tools. In addition to the opportunities for in silico analyses using expression data on the website and phenotypic data either collected by the user or extracted from archived datasets (e.g., GeneNetwork or the Mouse Phenome Project), the PhenoGen site offers numerous options for didactic, hands-on experiences with gene arrays in the realm of quantitative genetics for students and all levels of statistical scientists.

Acknowledgments This work was supported in part by NIH, NIAAA: R24AA013162 (BT), U01AA016663 (BT), U01AA016649 (PH), K01AA026922 (KK), and the Banbury Fund (BT).

References

Bergeson SE, Helms ML, O'Toole LA, Jarvis MW, Hain HS, Mogil JS, Belknap JK (2001) Quantitative trait loci influencing morphine antinociception in four mapping populations. Mamm Genome 12:546–553

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37:233–242

Hoffman PL, Bennett B, Saba LM, Bhave SV, Carosone-Link PJ, Hornbaker CK, Kechris KJ, Tabakoff B (2010) Using the Phenogen website for "in silico" analysis of morphine-induced analgesia: Identifying candidate genes. Addiction Biology doi: 10.1111/j.1369-1600.2010.00254.x

Hu W, Saba L, Kechris K, Bhave SV, Hoffman PL, Tabakoff B (2008) Genomic insights into acute alcohol tolerance. J Pharmacol Exp Ther 326:792–800

Lu L, Wei L, Peirce JL, Wang X, Zhou J, Homayouni R, Williams RW, Airey DC (2008) Using gene expression databases for classical trait QTL candidate gene discovery in the BXD recombinant inbred genetic reference population: mouse forebrain weight. BMC Genomics 9:444

Printz MP, Jirout M, Jaworski R, Alemayehu A, Kren V (2003) Genetic models in applied physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. J Appl Physiol 94:2510–2522

Saba L, Bhave SV, Grahame N, Bice P, Lapadat R, Belknap J, Hoffman PL, Tabakoff B (2006) Candidate genes and their regulatory elements: alcohol preference and tolerance. Mamm Genome 17:669–688

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717

Tabakoff B, Saba L, Kechris K, Hu W, Bhave SV, Finn DA, Grahame NJ, Hoffman PL (2008) The genomic determinants of alcohol preference in mice. Mamm Genome 19(5):352–365

Tabakoff B, Saba L, Printz M, Flodman P, Hodgkinson C, Goldman D, Koob G, McBride W, Bell G, Hübner N, Heinig M, Mangion J, LeGault L, Dongier M, Conigrave K, Whitfield J, Saunders J, Grant B, Hoffman PL, Health World, Organization/International Society for Biomedical Research on Alcoholism Study on State, Trait Markers of Alcohol Use, Dependence Investigators (2009)

Genetical genomic determinants of alcohol consumption in rats and humans. BMC Biol 7:70. doi:10.1186/1741-7007-7-70

Williams RW, Bennett B, Lu L, Gu J, DeFries JC, Carosone-Link PJ, Rikke BA, Belknap JK, Johnson TE (2004) Genetic structure of the LXS panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis. Mamm Genome 15(8): 637–647